Coffee & Health

How daily coffee routine affects
the our health firures

Maria / Pavel / Eetu / Liu (Suphitcha)

GROUP 6

Let’s be honest: as programming students, we’re practically made of coffee at this point—there’s probably more espresso than blood in circulation during exam weeks. Coffee isn’t just a beverage; it’s borderline academic fuel. On top of that, the university’s 80-cent coffee deal has turned caffeine into a daily ritual for both students and lecturers. So beyond pure academic curiosity, there’s a personal and almost professional interest in understanding what all this coffee is actually doing to our bodies. If cheap daily caffeine is part of campus culture, it’s fair to ask: is it helping us function—or quietly wrecking our sleep and stress levels?

Motivation

Objectives

1. Determine a Caffeine Threshold:
Can we identify a daily Caffeine mg threshold above which the probability of affecting Sleeping Quality or Stress Level significantly increases, holding other factors constant?

2. Identify Key Health Predictors:
Which lifestyle habits Coffee-, Alcohol Consumption, Smoking, or Physical Activity Hours — are the most significant predictors of Health Issues?

01

Data

What we are working with?

Global Coffee Health Dataset

Short discription:
The GlobalCoffeeHealth dataset contains 10,000 synthetic records reflecting real-world patterns of coffee consumption, sleep behavior, and health outcomes across 20 countries.

Column	Type	Description
ID	Integer	Unique record ID (1–10000)
Age	Integer	Age of participant (18–80 years)
Gender	Categorical	Male, Female, Other
Country	Categorical	Country of residence (20 countries)
Coffee_Intake	Float	Daily coffee consumption in cups (0–10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
BMI	Float	Body Mass Index (15–40)
Heart_Rate	Integer	Resting heart rate (50–110 bpm)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Physical_Activity_Hours	Float	Weekly physical activity (0–15 hours)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)
Occupation	Categorical	Office, Healthcare, Student, Service, Other
Smoking	Boolean	0 = No, 1 = Yes
Alcohol_Consumption	Boolean	0 = No, 1 = Yes

Source:
Kaggle | https://www.kaggle.com/datasets/uom190346a/global-coffee-health-dataset | Public Domain

Column	Type	Description
ID	Integer	Unique record ID (1–10000)
Age	Integer	Age of participant (18–80 years)
Gender	Categorical	Male, Female, Other
Country	Categorical	Country of residence (20 countries)
Coffee_Intake	Float	Daily coffee consumption in cups (0–10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
BMI	Float	Body Mass Index (15–40)
Heart_Rate	Integer	Resting heart rate (50–110 bpm)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Physical_Activity_Hours	Float	Weekly physical activity (0–15 hours)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)
Occupation	Categorical	Office, Healthcare, Student, Service, Other
Smoking	Boolean	0 = No, 1 = Yes
Alcohol_Consumption	Boolean	0 = No, 1 = Yes

Column	Type	Description
ID	Integer	Unique record ID (1–10000)
Age	Integer	Age of participant (18–80 years)
Gender	Categorical	Male, Female, Other
Country	Categorical	Country of residence (20 countries)
Coffee_Intake	Float	Daily coffee consumption in cups (0-10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
BMI	Float	Body Mass Index (15–40)
Heart_Rate	Integer	Resting heart rate (50–110 bpm)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Physical_Activity_Hours	Float	Weekly physical activity (0–15 hours)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)
Occupation	Categorical	Office, Healthcare, Student, Service, Other
Smoking	Boolean	0 = No, 1 = Yes
Alcohol_Consumption	Boolean	0 = No, 1 = Yes

Column	Type	Description
ID	Integer	Unique record ID (1–10000)
Age	Integer	Age of participant (18–80 years)
Gender	Categorical	Male, Female, Other
Country	Categorical	Country of residence (20 countries)
Coffee_Intake	Float	Daily coffee consumption in cups (0–10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
BMI	Float	Body Mass Index (15–40)
Heart_Rate	Integer	Resting heart rate (50–110 bpm)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Physical_Activity_Hours	Float	Weekly physical activity (0–15 hours)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)
Occupation	Categorical	Office, Healthcare, Student, Service, Other
Smoking	Boolean	0 = No, 1 = Yes
Alcohol_Consumption	Boolean	0 = No, 1 = Yes

Column	Type	Description
ID	Integer	Unique record ID (1–10000)
Age	Integer	Age of participant (18–80 years)
Gender	Categorical	Male, Female, Other
Country	Categorical	Country of residence (20 countries)
Coffee_Intake	Float	Daily coffee consumption in cups (0–10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
BMI	Float	Body Mass Index (15–40)
Heart_Rate	Integer	Resting heart rate (50–110 bpm)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Physical_Activity_Hours	Float	Weekly physical activity (0–15 hours)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)
Occupation	Categorical	Office, Healthcare, Student, Service, Other
Smoking	Boolean	0 = No, 1 = Yes
Alcohol_Consumption	Boolean	0 = No, 1 = Yes

Column	Type	Description
ID	Integer	Unique record ID (1–10000)
Age	Integer	Age of participant (18–80 years)
Gender	Categorical	Male, Female, Other
Country	Categorical	Country of residence (20 countries)
Coffee_Intake	Float	Daily coffee consumption in cups (0–10)
Caffeine_mg	Float	Estimated daily caffeine intake in mg (1 cup ≈ 95 mg)
Sleep_Hours	Float	Average hours of sleep per night (3–10 hours)
Sleep_Quality	Categorical	Poor, Fair, Good, Excellent (based on sleep hours)
BMI	Float	Body Mass Index (15–40)
Heart_Rate	Integer	Resting heart rate (50–110 bpm)
Stress_Level	Categorical	Low, Medium, High (based on sleep hours and lifestyle)
Physical_Activity_Hours	Float	Weekly physical activity (0–15 hours)
Health_Issues	Categorical	None, Mild, Moderate, Severe (based on age, BMI, and sleep)
Occupation	Categorical	Office, Healthcare, Student, Service, Other

Smoking	10000.0	0.20040	0.400320	0.0	0.00	0.0	0.000	1.0
Alcohol_Consumption	10000.0	0.30070	0.458585	0.0	0.00	0.0	1.000	1.0

01.5

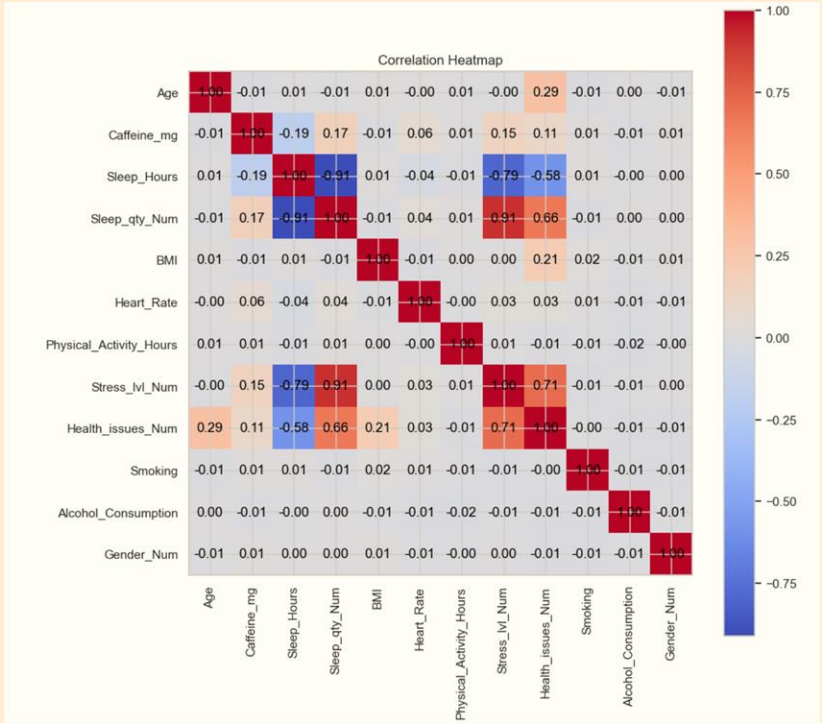
Data understanding

Correlations

```
# Create mapping dictionaries
stress_map = {'Low': 0, 'Medium': 1, 'High': 2}
sleep_map = {'Excellent': 0, 'Good': 1, 'Fair': 2, 'Poor': 3}
health_map = {'None': 0, 'Mild': 1, 'Moderate': 2, 'Severe': 3}
gender_map = {'Female': 0, 'Male': 1, 'Other': np.nan}
```

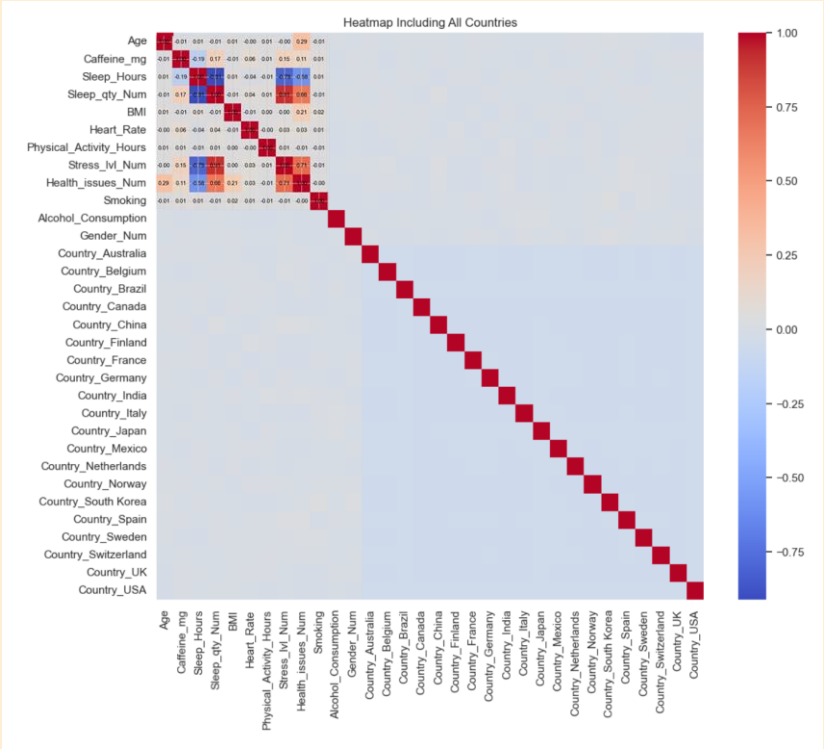
Remap the most essential features
to create the correlation matrix

Are there any correlations?



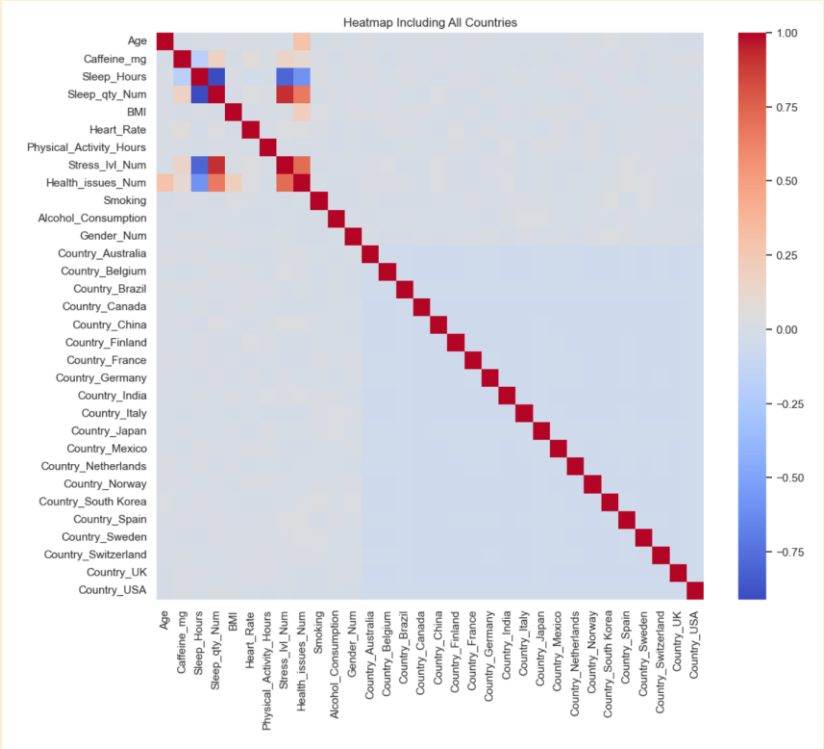

```
# Apply the mapping to create new numerical columns
df['Stress_lvl_Num'] = df['Stress_Level'].map(stress_map)
df['Sleep_qty_Num'] = df['Sleep_Quality'].map(sleep_map)
df['Health_issues_Num'] = df['Health_Issues'].map(health_map)
df['Gender_Num'] = df['Gender'].map(gender_map)
```

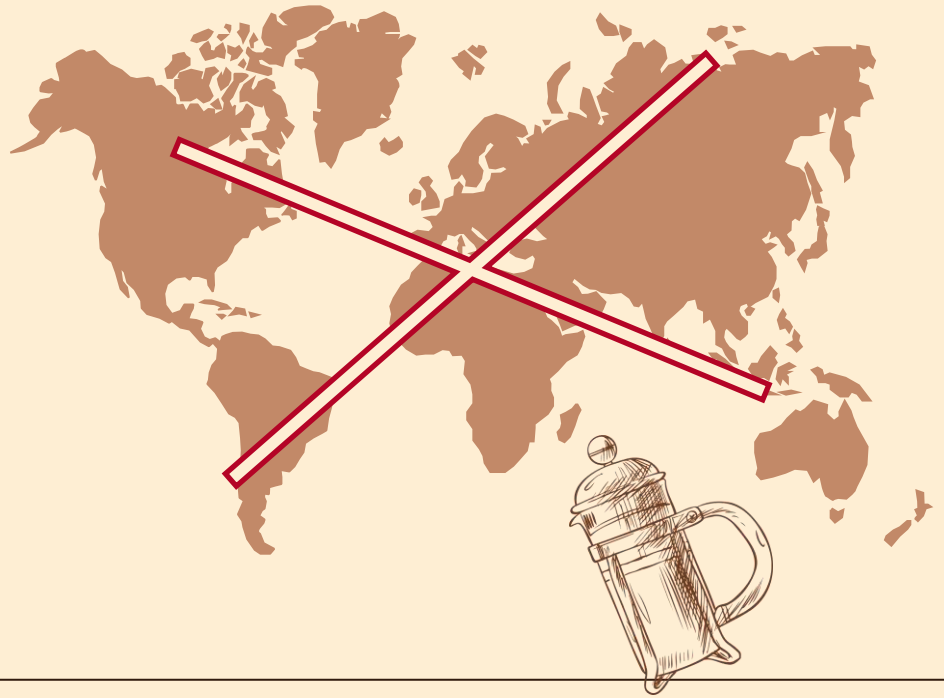
Is there any location based correlations ?



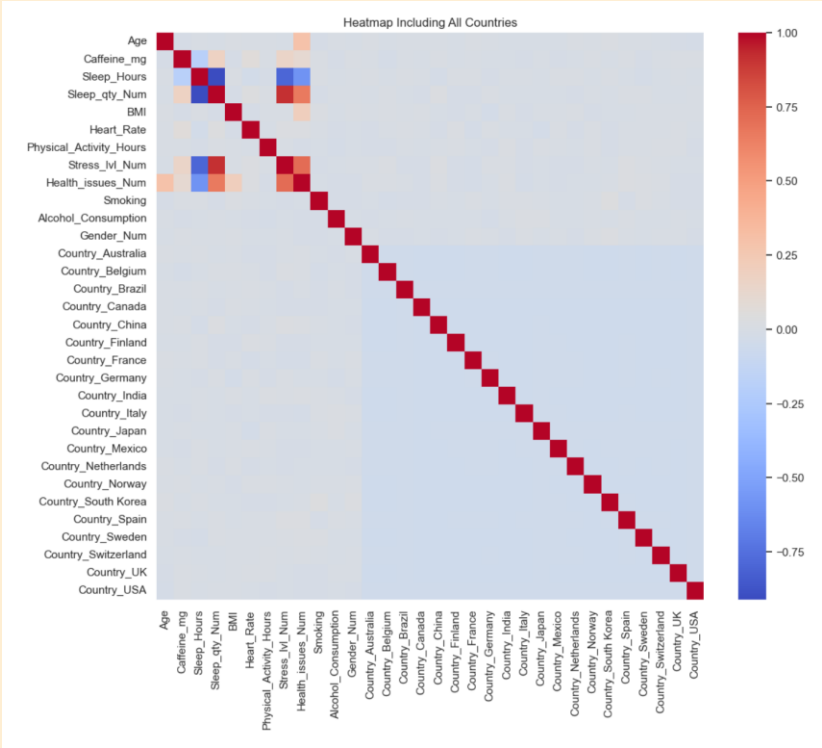
```
df_country_dummies = pd.get_dummies(df['Country'], prefix='Country')
df_merged = pd.concat([NumDf, df_country_dummies], axis=1)

plt.figure(figsize=(12, 10))
sns.heatmap(df_merged.corr(), cmap='coolwarm')
```

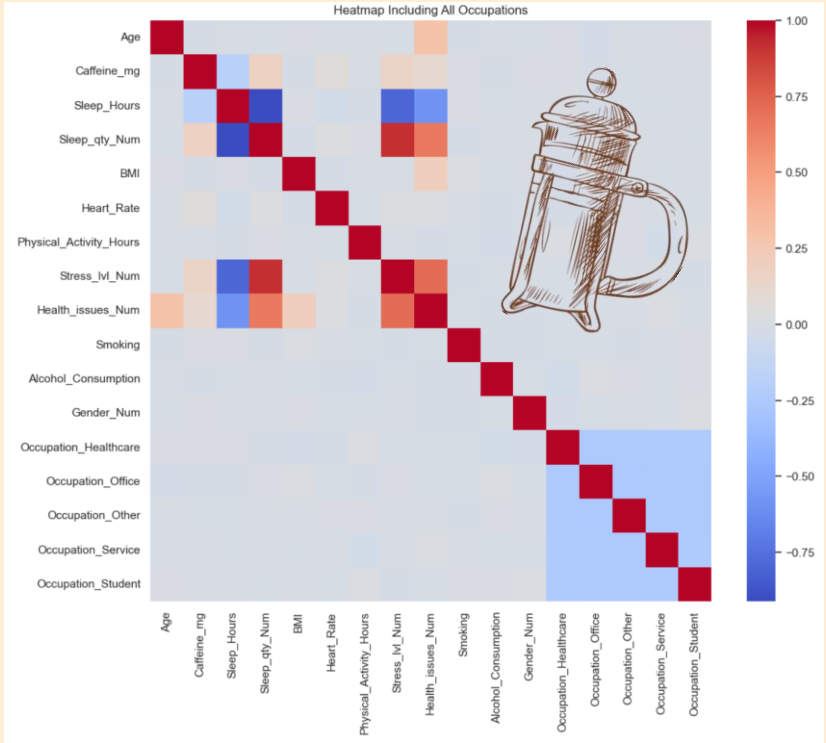
Not so much


Heatmap Including All Countries
Age
Caffeine_mg
Sleep_Hours
Sleep_qty_Num
BMI
Heart_Rate
Physical_Activity_Hours
Stress_lvl_Num
Health_issues_Num
Smoking
Alcohol_Consumption
Gender_Num
Country_Australia
Country_Belgium
Country_Brazil
Country_Canada
Country_China
Country_Finland
Country_France
Country_Germany
Country_India
Country_Italy
Country_Japan
Country_Mexico
Country_Netherlands
Country_Norway
Country_South Korea
Country_Spain
Country_Sweden
Country_Switzerland
Country_UK
Country_USA
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

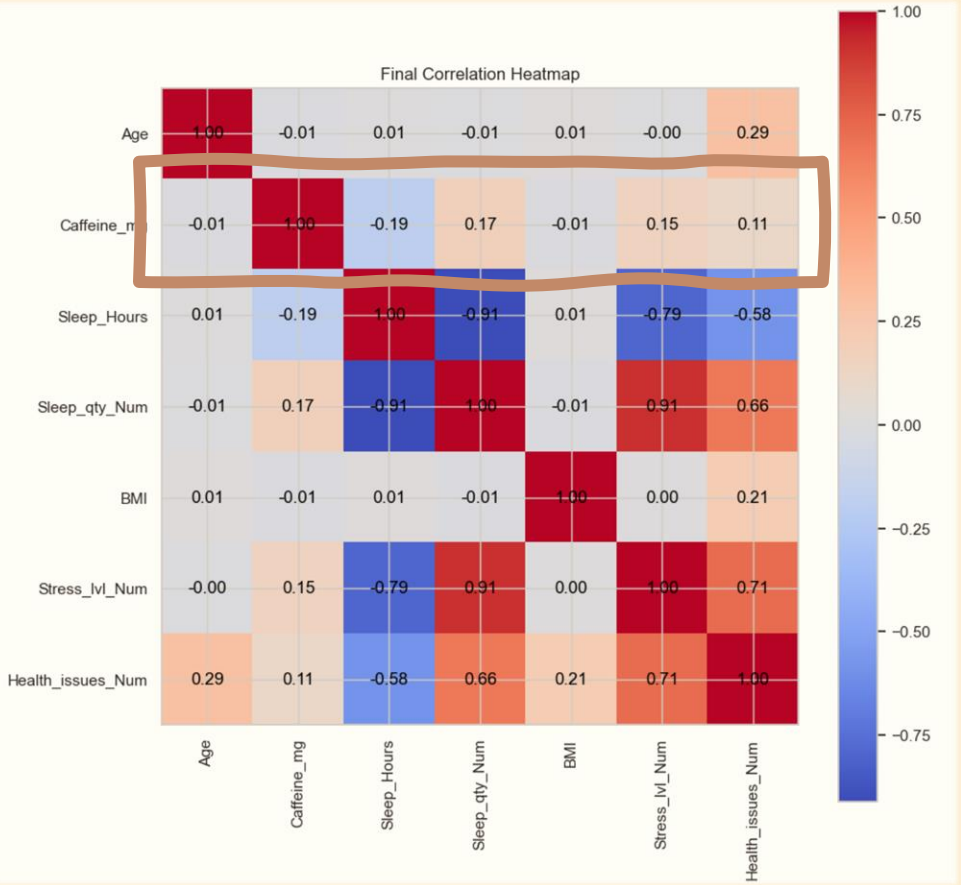


Same for occupation

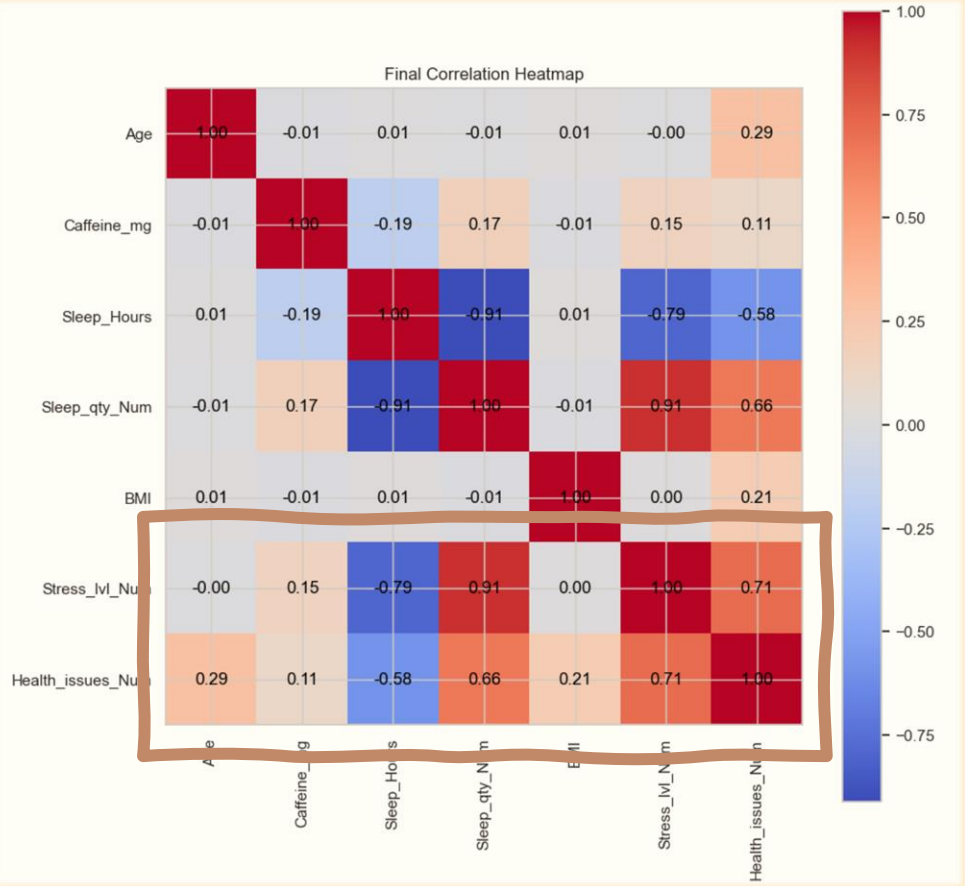




In total


Final Correlation Heatmap
Age
Caffeine_mg
Sleep_Hours
Sleep_qty_Num
BMI
Stress_lvl_Num
Health_issues_Num
1.00 -0.01 0.01 -0.01 0.01 -0.00 0.29
-0.01 1.00 -0.19 0.17 -0.01 0.15 0.11
0.01 -0.19 1.00 -0.91 0.01 -0.79 -0.58
-0.01 0.17 -0.91 1.00 -0.01 0.91 0.66
0.01 -0.01 0.01 -0.01 1.00 0.00 0.21
-0.00 0.15 -0.79 0.91 0.00 1.00 0.71
0.29 0.11 -0.58 0.66 0.21 0.71 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

In total


Final Correlation Heatmap
Age
Caffeine_mg
Sleep_Hours
Sleep_qty_Num
BMI
Stress_lvl_Num
Health_issues_Num
1.00 -0.01 0.01 -0.01 0.01 -0.00 0.29
-0.01 1.00 -0.19 0.17 -0.01 0.15 0.11
0.01 -0.19 1.00 -0.91 0.01 -0.79 -0.58
-0.01 0.17 -0.91 1.00 -0.01 0.91 0.66
0.01 -0.01 0.01 -0.01 1.00 0.00 0.21
-0.00 0.15 -0.79 0.91 0.00 1.00 0.71
0.29 0.11 -0.58 0.66 0.21 0.71 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

02

Data preparation

specific binary outcomes

```
# Target 1: Is the person experiencing high stress?
df['Is_High_Stress'] = (df['Stress_lvl_Num'] >= 1).astype(int) # Medium
# Target 2: Is the person experiencing poor sleep?
df['Is_Poor_Sleep'] = (df['Sleep_qty_Num'] >= 2).astype(int) # Fair
# Target 3: Does the person have any health issues ?
df['Has_Health_Issues'] = (df['Health_issues_Num'] >= 1).astype(int) # Mild
```

"middle ground" approach with the lowest possible values

03

Modeling

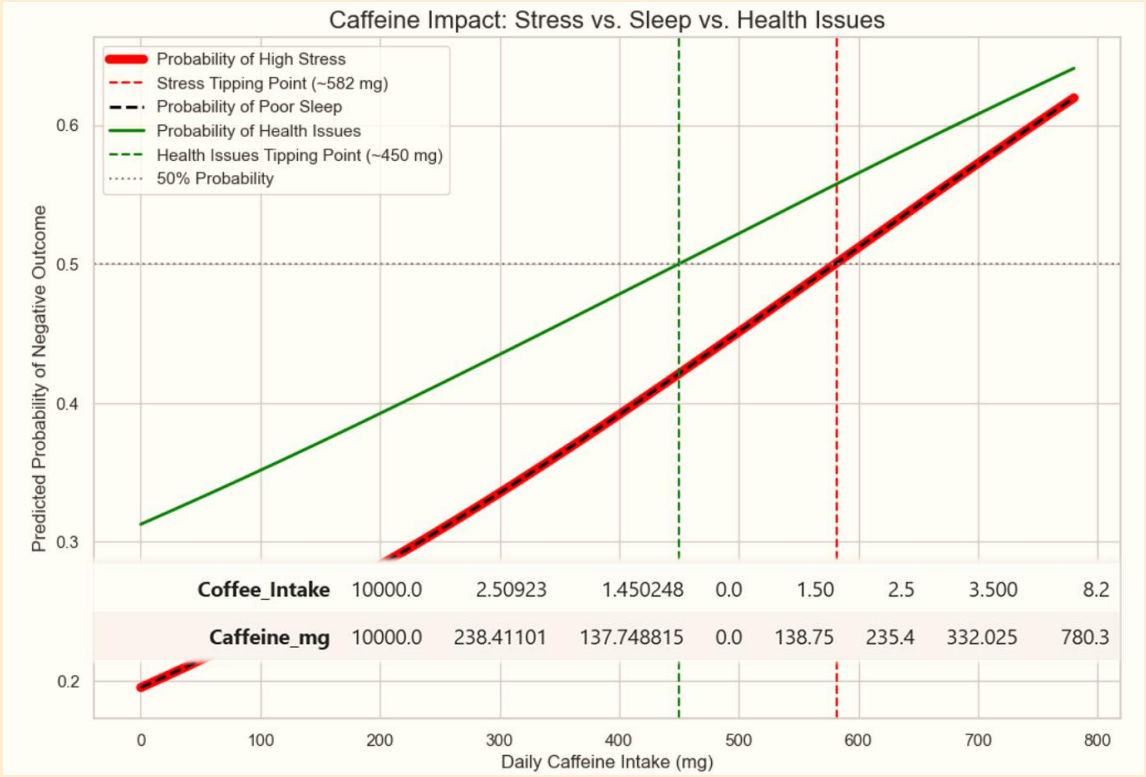
Thresholds and predictions

Coffee effect on the health

We didn’t have the ability to check individual coffee tolerance and its effect on each person.

```python
# List of target variables and their descriptive labels
targets = [
    {'column': 'Is_High_Stress', 'label': 'High Stress'},
    {'column': 'Is_Poor_Sleep', 'label': 'Poor Sleep'},
    {'column': 'Has_Health_Issues', 'label': 'Health Issues'}
]

# Dictionary to store results for plotting
plot_data = {}

# --- Main loop for processing ---
for target in targets:
    print(f"--- Processing: {target['label']} ---")

    # 1. Select the feature (target variable)
    y = df[target['column']]

    # 2. Split data for the current model
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=randomstate)

    # 3. Train the model
    log_reg = LogisticRegression(random_state=randomstate)
    log_reg.fit(X_train, y_train)

    # 4. Predict probabilities on the caffeine range
    probs = log_reg.predict_proba(caffeine_range_df)[:, 1]

    # Store probabilities for later plotting
    plot_data[target['label']] = {'probs': probs}

    # 5. Find and print the highest probability
    max_prob = probs.max()
    print(f"Maximum predicted probability for {target['label']}: {max_prob:.2%}")
```

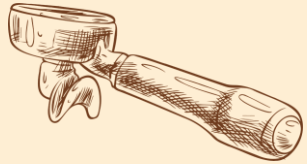
```
plt.legend()
plt.grid(True)
plt.show()
```



Coffee_Intake	10000.0	2.50923	1.450248	0.0	1.50	2.5	3.500	8.2
Caffeine_mg	10000.0	238.41101	137.748815	0.0	138.75	235.4	332.025	780.3

2-3x

Analysis Medium

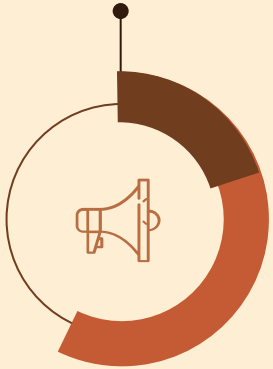


35%

Health

Moderate problems

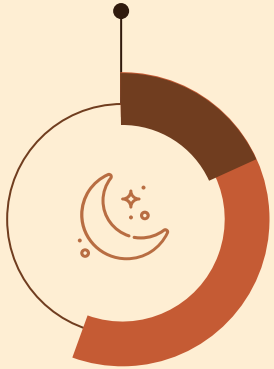
40%



Stress

Medium

40%



Sleep

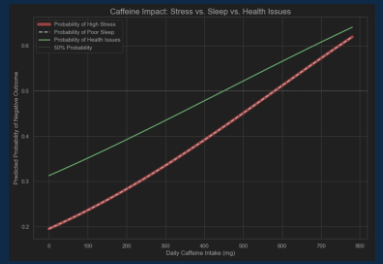
Fair

0.65 - 0.3 = 0.35 | 0.6 - 0.2 = 0.4

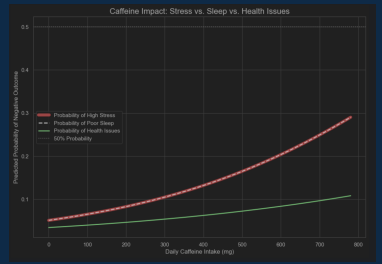
2-3x

Behind the scenes

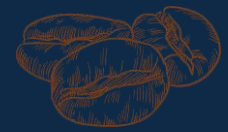
```
# --- Create three distinct binary targets ---
# Target 1: Is the person experiencing high stress?
df['Is_High_Stress'] = (df['Stress_Num'] >= 1).astype(int)
# Target 2: Is the person experiencing poor sleep?
df['Is_Poor_Sleep'] = (df['Sleep_Num'] >= 2).astype(int)
# Target 3: Does the person have any health issues ?
df['Has_Health_Issues'] = (df['Health_Num'] >= 1).astype(int)
```

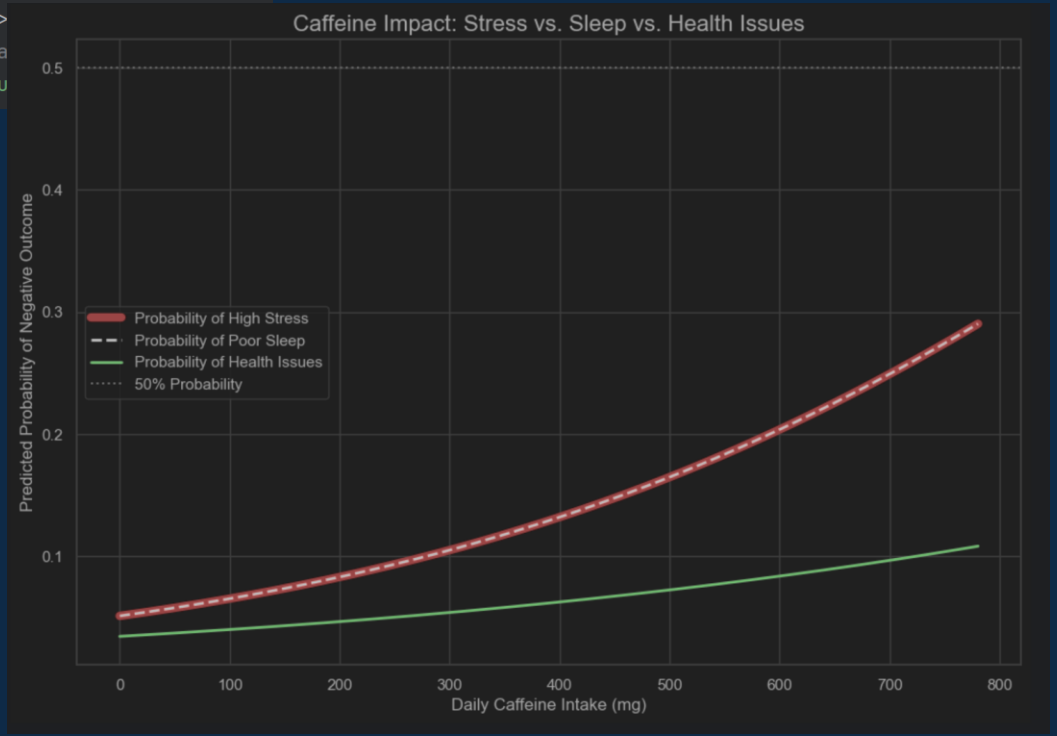



```
# --- Create three distinct binary targets ---
# Target 1: Is the person experiencing high stress?
df['Is_High_Stress'] = (df['Stress_Num'] >= 2).astype(int)
# Target 2: Is the person experiencing poor sleep?
df['Is_Poor_Sleep'] = (df['Sleep_Num'] >= 3).astype(int)
# Target 3: Does the person have any health issues ?
df['Has_Health_Issues'] = (df['Health_Num'] >= 2).astype(int)
```



```
# --- Create three distinct binary targets ---
# Target 1: Is the person experiencing high stress?
df['Is_High_Stress'] = (df['Stress_Num'] >= 2).astype(int)
# Target 2: Is the person experiencing poor sleep?
df['Is_Poor_Sleep'] = (df['Sleep_Num'] >
# Target 3: Does the person have any hea
df['Has_Health_Issues'] = (df['Health_Nu
```





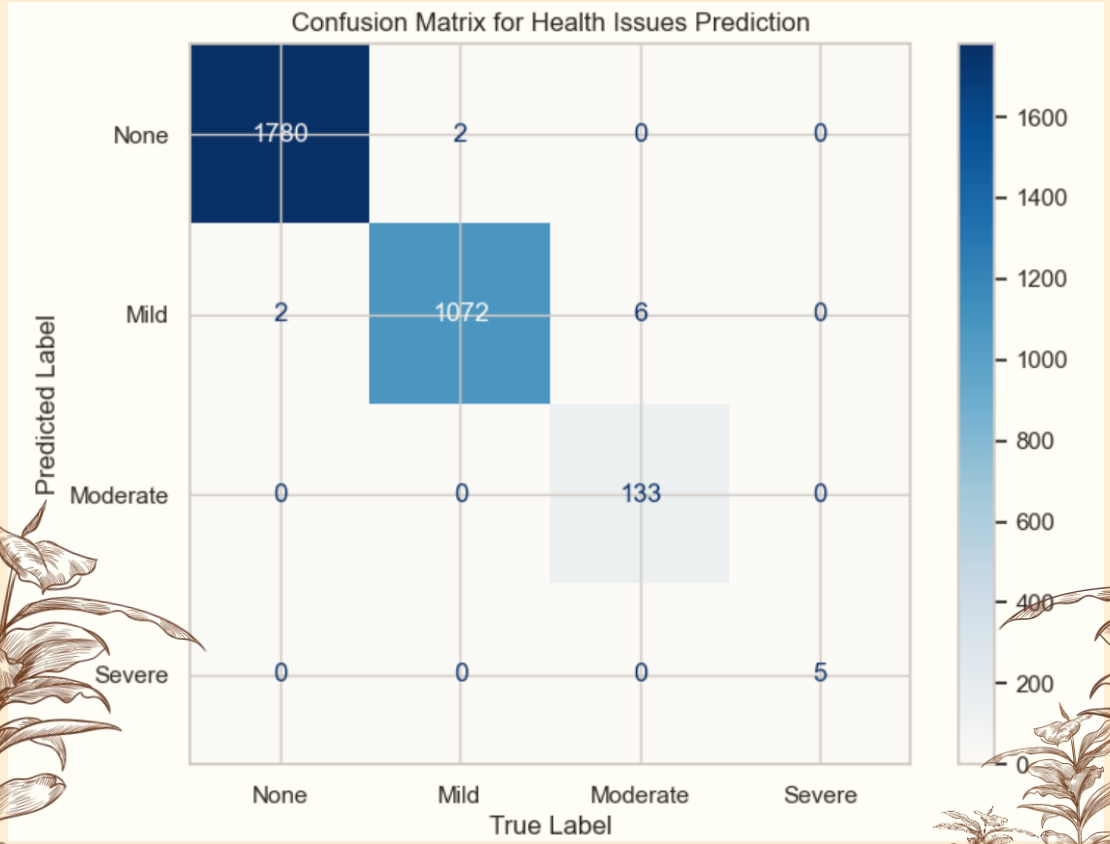
03.5

What affects health?

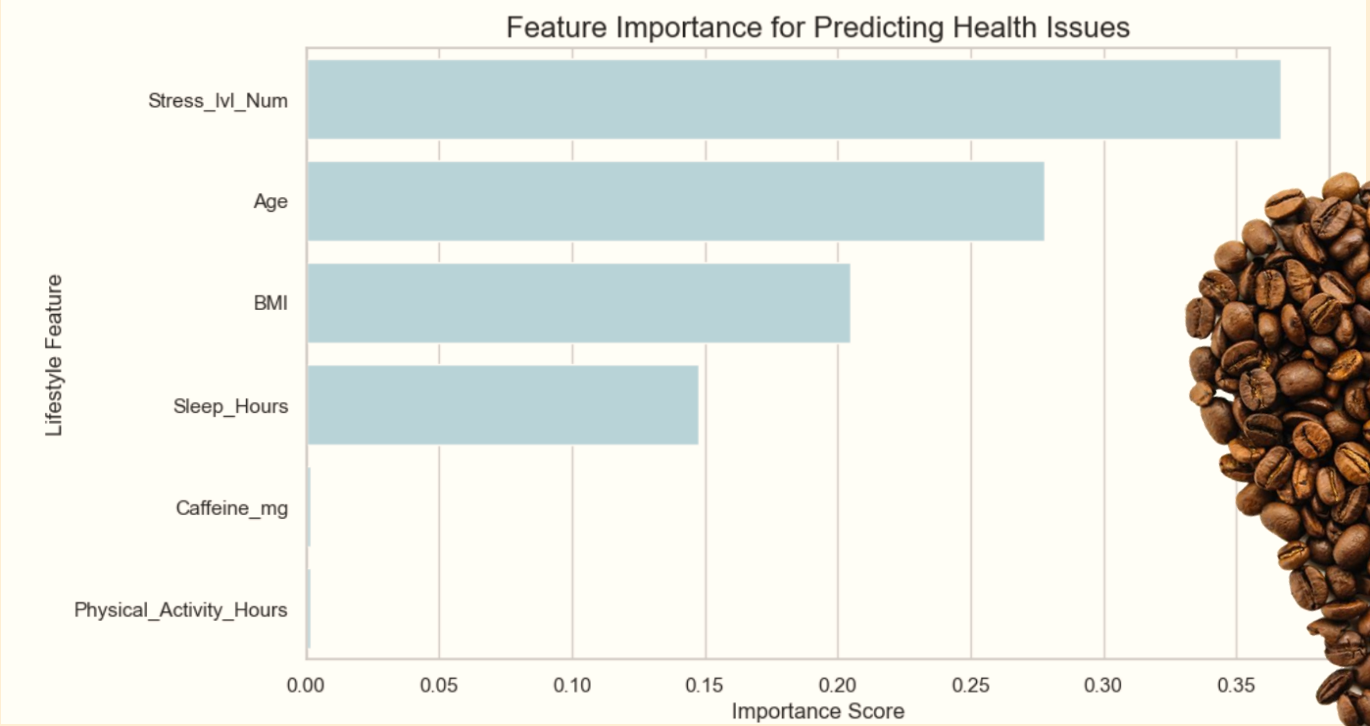
```
# Define Features (X) and Target (y)
X = df[[
  'Caffeine_mg',
  'Sleep_Hours',
  'Physical_Activity_Hours',
  'Stress_lvl_Num',
  'Age',
  'BMI'
]]
y = df['Health_Issues']
```

```
# Carry out 10-fold cross-validation
cv_scores = cross_val_score(rf_classifier, X, y, cv=10)
print(f"Average Accuracy: {cv_scores.mean():.2%} (+/- {cv_scores.std():.2%})")
```

Average Accuracy: 99.81% (+/- 0.17%)


Confusion Matrix for Health Issues Prediction
None
Mild
Moderate
Severe
Predicted Label
1780
2
0
0
2
1072
6
0
0
0
133
0
0
0
0
5
None
Mild
Moderate
Severe
True Label
1600
1400
1200
1000
800
600
400
200
0

Effect on the health

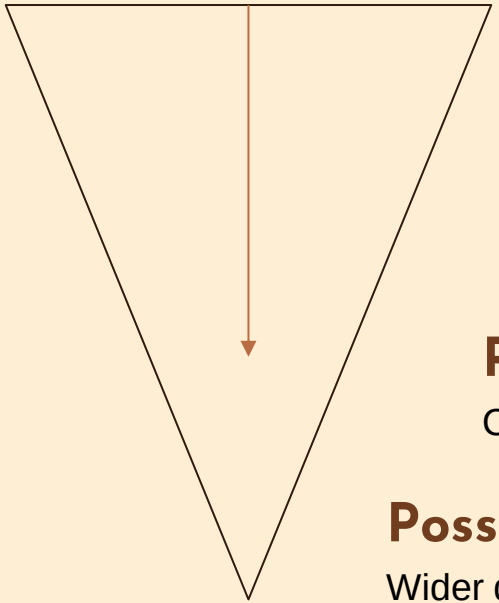


04

Evaluation

What have we learned

Our achievements / findings



Caffeine Treshold

Is secondary compared to other factors.

Health Predictions

High-precision prediction of health indicators.

Possible implementation

Concept of personal health assistant

Possible improvements

Wider dataset with timeline-based tracking

05

Deployment

Thresholds and predictions

05

Deployment

A potential outcome of this work is an application capable of tracking users' habits and providing personalized health feedback. However, to ensure its effectiveness, a wider dataset is required to evaluate the comparative impact of alcohol and cigarette use, as well as physical activity. Understanding these factors is essential to validate our findings and develop a safer, more reliable system.